coronasurveys.org

C’est quoi ce projet?

Nous savons que pour gérer au mieux la pandémie de Covid-19, les gouvernements ont besoin d’avoir les données les plus précises sur la situation. Or à l’heure actuelle, les seules données dont on dispose reposent sur les tests élaborés en laboratoire et sur le nombre d’appels traités par téléphone. Par conséquent, un grand nombre de cas non dépistés ne sont pas comptabilisés, comme ceux sans gravité ou guéris suffisamment vite pour ne pas être déclarés.

Notre étude porte sur un sondage – anonyme et très court (https://coronasurveys.org https://tinyurl.com/coronasurveysfrance) – au quel on demande de répondre quotidiennement. Anonyme car l’un des objectifs de ce sondage était de respecter scrupuleusement le RGPD (Règlement général sur la protection des données) et donc aucune donnée personnelle n’est demandée. Très court puisqu’il ne comporte que trois questions : (i) votre département ; (ii) le nombre de personnes dans le département sélectionné dont vous êtes susceptibles de connaître l’état de santé ; et (iii) le nombre de personnes à votre connaissance, ayant ou ayant eu des symptômes compatibles avec le COVID-19.

Pourquoi si peu de questions?

L’utilisation de ces trois questions présente deux avantages. Premièrement, ils ne demandent aucune donnée personnelle car nous visons à protéger la vie privée du participant. Deuxièmement le fait de récolter des information sur l'état de santé des contactes de participants permets d'élargir le nombre de personnes observée par le sondage, permettant d’avoir des estimations raisonnables même avec des taux de réponse relativement faibles. Le faible nombre de questions posées vise aussi à simplifier le processus de réponse qui ne demande que quelques seconds. Nous espérons ainsi accroître la participation.

Mais, est-ce que vous avez assez d’informations?

Cependant, le manque d'informations détaillées sur les participants rend le processus d'estimation difficile. Nous ne contrôlons pas la diffusion de l'enquête ni la couverture adéquate des régions, des groupes d'âge et d'autres paramètres. Il est quelque peu surprenant que, malgré ces limites, nous puissions obtenir une estimation approximative et constater qu'elle n'est pas très éloignée de celles obtenues avec d'autres techniques. Comme déjà souligné, l'une des raisons peut être que chaque participant rend compte de l'état de santé d'un grand échantillon (des centaines), ce qui augmente considérablement la

couverture de l'enquête. En résumant, les avantages les plus évidents de cette approche sont qu'elle est très simple à déployer et qu'elle peut donner des résultats très rapidement.

Comment ça se déroule donc le processus d'estimation?

Le processus pour obtenir l'estimation des cas est le suivant. Les réponses à l'enquête sont nettoyées en identifiant et en éliminant les valeurs aberrantes. Premièrement, nous éliminons les réponses inhabituellement grandes en termes de nombre de personnes connues. Plus précisément, nous supprimons les entrées qui dépassent le quartile supérieur (Q3) de plus de 1,5 fois l'intervalle interquartile (Q3-Q1). Deuxièmement nous éliminons les réponses qui présentent une proportion trop élevé de personnes symptomatiques déclarées (plus de 30 %). Ces derniers sont supprimés parce que notre objectif est de sonder la population générale qui n'est pas particulièrement en contact avec des cas symptomatiques plutôt que des soignants qui pourraient effectivement avoir un ratio bien supérieur à 30%.

Une fois les données épurées, nous les utilisons pour obtenir le pourcentage de cas dans la population qui est connu des personnes interrogées. Ensuite, nous extrapolons naïvement ce ratio à l'ensemble de la population du pays. Plus formellement, supposons que nous avons *n* réponses à l'enquête, et que la *i*-ème réponse aille une valeur de portée, r_i (reach), pour la première question, et une valeur de cas, c_i, pour la deuxième question. Prenons ensuite, *d = 150* représentant le nombre de Dunbar [2, 16], c'est-à-dire le nombre attendu de personnes avec lesquelles on entretient des relations sociales stables. Nous obtenons les estimations suivantes pour une zone géographique avec une population *P*.

$$E_w = P \cdot \frac{\sum_i c_i}{\sum_i r_i}$$

$$E_m = P \cdot \frac{\sum_i (c_i / r_i)}{n}$$

$$E_d = P \cdot \frac{\sum_i c_i}{nd}$$

Etant donnée la proportion d'infection associée à chaque réponse c_i/r_i, la première estimation E_w pèse l'impacte de chaque proportion par sa valeur de portée, r_i , alors que la deuxième E_m représente une moyenne simple des proportions c_i/r_i. Nous avons également observé qu'il est souvent difficile pour les utilisateurs d'estimer le nombre de personnes qu'ils connaissent. Avec l'estimation E_d nous ignorons donc la valeur de chaque portée déclarée r_i et nous la remplaçons par le nombre de Dunbar *d*.

J’ai vu que les graphes montrent des intervals de confiance; comment sont-ils calculées?

En supposant pour l'instant une indépendance entre les observations, nous pouvons estimer l'incertitude associée aux estimations ci-dessus et en dériver des intervalles de confiance. Pour cela, nous modélisons chaque réponse comme une petite enquête pour obtenir un ensemble plus large d'observations. En particulier, dans le cas de E_w et E_m, chaque “petite-enquête” comporte r_i observations, donnant un total de $r = \sum_{i=1}^{n} r_i$. Dans le cas particulier de E_d, nous avons $r = nd$.

Étant donné que nous travaillons avec des observations binaires (absence ou présence de symptômes), nous considérons une distribution binomiale, *Bin(r,p)*, où *p* correspond à la probabilité d'avoir de symptômes de la maladie. Cela donne les estimations ponctuelles suivantes pour p dans les trois cas.

$$\hat{p}_w = \frac{\sum_i c_i}{\sum_i r_i}$$
$$\hat{p}_m = \frac{\sum_i (c_i / r_i)}{n}$$
$$\hat{p}_d = \frac{\sum_i c_i}{nd}$$

En appliquant le théorème central limite, nous obtenons l’interval de confiance suivante:

$$\hat{p} \in \{\hat{p}_w, \hat{p}_m, \hat{p}_d\} \text{ as } \widehat{p} \pm z\sqrt{\frac{\widehat{p}(1-\widehat{p})}{r}}$$

Pour une confiance à 95 %, nous prenons donc z = 1,96 (le quantile correspondant de la distribution normale standard).